A randomized model ensemble approach for reconstructing signals from faulty sensors

Piero Baraldi, Giulio Gola, Enrico Zio, Davide Roverso, M. Hoffmann

A randomized model ensemble approach for reconstructing signals from faulty sensors

P. Baraldi[1], G. Gola[2,*], E. Zio[1], D. Roverso[2], M. Hoffmann[2]

[1]Department of Energy, Polytechnic of Milan, Via Ponzio 34/3, 20133, Milano, Italy
[2]Institute for Energy Technology, Os Aleè 5, 1767, Halden, Norway
[*]corresponding author: giulio.gola@hrp.no; tel: +47 69212215

Abstract

On-line sensor monitoring aims at detecting anomalies in sensors and reconstructing their correct signals during operation. The techniques used for signal reconstruction are commonly based on auto-associative regression models. In full scale implementations however, the number of sensors to be monitored is often too large to be handled effectively by a single reconstruction model. In this paper we propose to tackle the problem by resorting to a pool (ensemble) of reconstruction models, each one handling an individual group of signals. This approach involves two main technical steps: firstly, a procedure for constructing signal groups, and secondly a procedure for combining the outputs of the reconstruction models associated to the groups. For the signal grouping step, a wrapper optimization search is proposed to identify the optimal number of groups in the ensemble and the size of the groups. For the model output aggregation step, a simple arithmetic average is adopted. Ensemble accuracy and robustness is achieved by promoting diversity between the signal groups through the use of the Random Feature Selection Ensemble (RFSE) technique in combination with the Bootstrapping AGGregatING (BAGGING) technique for training data selection. The individual reconstruction models are based on Principal Components Analysis (PCA). The proposed approach has been applied to a real case study concerning 215 signals monitored at a Finnish nuclear pressurized water reactor. The results obtained have been compared with those achieved by an equivalent ensemble of models based on a grouping directly optimized by a Multi-Objective Genetic Algorithm (MOGA).

Keywords: Sensor monitoring, Signal reconstruction, Ensemble, Diversity, Random feature selection, Bagging

1. Introduction

Sensors contribute to the safe and efficient operation of modern plants by conveying information on the plant state to the automated controls and the operators. To avoid misleading information which may lead to unsafe and/or inefficient states of operation, it is important to detect sensor malfunctions and possibly reconstruct the incorrect signals. This requires monitoring the sensor performance and has the potential benefits of reducing unnecessary sensor maintenance and increasing confidence in the recorded values of the monitored parameters, with important consequences on system operation, production and accident management [1, 2].

The problem of validating the signals recorded by sensors can be tackled by means of empirical models based on fuzzy logic [3, 4] and neural networks [5, 6]. In particular, auto-associative models have been applied to the validation of nuclear signals [7-9]. Nevertheless, the single-model approaches typically used for signal validation can only handle a limited number of signals whereas practical applications often deal with a very large number of signals [1, 2].

In this work, a procedure is proposed for the reconstruction of signals coming from faulty sensors among a large set. The procedure is tailored for realistic applications where the number of measured signals is too large to be handled effectively by a single reconstruction model [2, 10-12]. The approach is based on the subdivision of the set of signals into overlapping groups, the development of a reconstruction model for each group of signals and the combination of the outcomes of the models within an ensemble approach [13-22] (Figure 1).

An additional advantage of adopting ensembles of diverse models is an increased robustness of the ensemble-aggregated output [17, 21-26]. Indeed, the conjecture is that, when performing the ensemble signal reconstruction, if the signal predictions obtained by the individual models are diverse (e.g. the reconstruction errors are different in magnitude and sign), their opportune aggregation will provide a more accurate and robust signal reconstruction [23, 24, 26]. Theoretical studies have investigated the concept of diversity amongst the models of an ensemble [20-23, 25, 26] and the ways of appropriately aggregating the outcomes of the diverse models [19, 22, 25].

This work intends to contribute a practical application of theoretical concepts of empirical ensemble modeling to a large-scale problem of monitoring and reconstructing a large amount of signals recorded by sensors at a nuclear power plant.

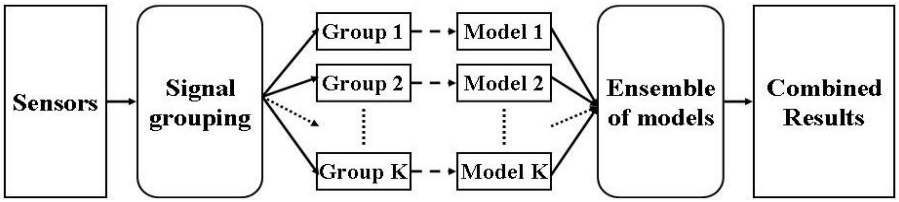

Figure 1. The multi-group ensemble approach to signal reconstruction

Two issues are central to the ensemble approach: (1) the construction of the signal groups and (2) the combination of the outcomes of the individual models developed on the basis of the groups.

In practice, the aspects most relevant to the grouping of signals are:

(i) the groups size: groups made of a reasonably small number of signals are both more accurate and easier to handle [10-12, 16-18];

(ii) the ensemble size: limiting the number of groups in the ensemble helps reducing the computational cost;

(iii) the diversity of the groups and of the associated models: ensembles based on diverse models are generally more reliable and robust [17, 19, 21, 23-26].

In this work, the groups and ensemble sizes are derived by means of a wrapper optimization. To promote group diversity, signals groups are randomly generated resorting to the Random Feature Selection Ensemble (RFSE) technique [27]. The groups thereby created are then used to develop a corresponding number of signal reconstruction Principal Component Analysis (PCA) [28-31] models. The models are then trained on different data sets randomly generated using the Bootstrapping AGGregatING (BAGGING) approach [23, 24, 27] in order to inject further diversity in the ensemble of models. The methods for both signal grouping and reconstruction modelling have been chosen simple, consistently with the idea that simplicity is an added value in nuclear safety, where possible. Other more sophisticated methods to generate different training data sets exist, e.g. AdaBoost [24], which seem worth consideration in future research for nuclear applications.

Regarding the integration of the outcomes of the individual models, many techniques can be adopted ranging from the simple arithmetic average to weighted average [17], local fusion [32] and dynamic integration techniques [19, 20], with increasing computational costs. In practice, there is no single combination rule that is universally recognized better than the others and it is still not clear that more sophisticated and complicated aggregation techniques are actually beneficial to the ensemble performance; then, for the same reason stated above and in absence of any other prior information, the simple arithmetic average is used as a valid choice for combining the models predictions [24, 33, 34].

The paper is organized as follows. Section 2 presents the problem of signal grouping in general terms. In Section 3, the wrapper randomized approach for generating diverse groups of signals is described in details. An application is illustrated in Section 4, with regards to a real case study concerning the reconstruction of a data set of 215 signals measured at a Finnish nuclear Pressurized Water Reactor (PWR) located in Loviisa. Two more case studies have been considered for verification: the first concerns 84 signals measured at a Swedish nuclear Boiling Water Reactor (BWR) situated in Oskarshamn, the second the reconstruction of 920 simulated signals of the Swedish Forsmark-3 Boiling Water Reactor (BWR). For the Loviisa case study, a comparison is made with a Multi-Objective Genetic Algorithm (MOGA) optimization procedure [35-38] for ensemble group signals selection developed in [17, 18]. A discussion on the advantages and limitations of the proposed procedure is offered in the last Section. Finally, in an attempt to make the paper self-consistent an Appendix reports a brief synthesis of the basic concepts of Principal Components Analysis for signal reconstruction.

2. Grouping signals for diversity and optimal ensemble performance

Given a set of available $n >> 1$ sensors' signals f_i, $i = 1,2,...,n$, signal grouping aims at constructing K groups of $m_k << n$ signals, $k = 1,2,...,K$, with given required characteristics.

The selection of the signals to insert in each group should be driven by both the individual properties of the groups and the global properties of the ensemble. Properties individually related to a group are, for example, the mutual information content of the signals in the group and the reconstruction performances of the associated model [11, 12, 16]; global properties of the groups ensemble are, for example, the diversity among the groups and a good redundancy of the signals in the ensemble, i.e., an adequate number of diverse groups containing a same signal [2, 10, 17, 18].

Heuristic methods such as Multi-Objective Genetic Algorithms (MOGAs) [38] have proved effective in scanning the large search space of possible groups[1] to optimize their individual properties, e.g. by maximizing the correlation of the signals in the groups [11, 16-18] and minimizing their reconstruction errors [12]. MOGA approaches have however shown some limitations in guaranteeing the mentioned global ensemble properties, e.g. diversity among the groups, adequate signal inclusion and redundancy, at the basis of the optimality and robustness of the performance of the ensemble [18]. Furthermore, the high computational cost required to run a MOGA search renders the method unfeasible for large-scale applications involving thousands of signals.

In this work, the RFSE technique is exploited to ensure good global ensemble properties of group diversity, signal coverage, and redundancy. The group size parameters, i.e. the number of signals in the groups and the number of groups in the ensemble, are determined by means of a wrapper search aimed at maximizing the ensemble performance in terms of minimum reconstruction error. PCA signal reconstruction models are trained on data sets constructed by BAGGING to inject further diversity in the models themselves.

3. The randomized wrapper approach to signal grouping

As previously stated, high diversity in the ensemble of models is beneficial to its performance of signal reconstruction. To this aim, in this work diversity is imposed onto the PCA models by randomizing the features of the groups upon which they are built with the RFSE technique [27] and the data upon which they are trained with the BAGGING technique [23, 24, 27]. Optimization of the group size m and ensemble size K, is also carried out to improve the performance of the ensemble.

3.1 Injecting and verifying (input) diversity in the signal groups by RFSE

The RFSE technique consists in randomly sampling, with replacement, from the n available signals K subsets $\mathbf{F}_k$, $k = 1, 2, ..., K$, each constituted by m signals. This guarantees high signal diversity in the groups upon which the PCA models are built and provides a much faster construction of the signal groups compared to the optimality-driven searches, e.g. MOGA-based [17, 18]). Indeed, RFSE is a completely random technique in which no optimization of the composition of the individual groups is sought, i.e. no relevance is given, for example, to the correlation between the signals in the groups (as in the *filter* MOGA search presented in [11, 16-18]) or to their capability of efficiently reconstructing one another (as in the *wrapper* MOGA search presented in [12]). The coverage of all the signals is not a priori guaranteed by the random nature of the RFSE approach itself; nevertheless, since the probability that a signal does not appear in any group is $\left(\frac{n-m}{n}\right)^K$, a reasonable choice of the values of the ensemble parameters m and K can in practice guarantee coverage of all the signals in the ensemble with adequate redundancy, as shown in Section 4.

As previously mentioned, an optimal signal grouping structure should ensure on one side a good signal redundancy in the groups (i.e., for any signal there is a suitable number of groups containing it) but, on the other side, a diverse composition of the groups in terms of the signals contained. In other words, the groups must partially overlap (in order to have each signal included in more than one group), while being sufficiently diverse among each other.

An empirical measure is here proposed to verify the diversity injected by the RFSE in the resulting ensemble grouping structure in terms of the diversity in the signal composition of the different groups (the so-called *input* diversity). Let us consider a generic ensemble of K groups with different sizes m_k, $k = 1, 2, ..., K$. The pairwise diversity between two generic groups k_1 and k_2 of sizes m_{k_1} and m_{k_2}, respectively, is computed as:

[1] Considering n signals, the number of possible groups of different sizes $m = 1, 2, ..., n$ that can be generated is equal to $\sum_{m=1}^{n} \frac{n!}{m!(n-m)!}$.

$$div_{in}^{k_1,k_2} = \frac{1}{1+\exp(12\beta_{com}^{k_1,k_2} - 6)} \tag{1}$$

where $\beta_{com}^{k_1,k_2} = n_{com}^{k_1,k_2} \Big/ \max\{m_{k_1}, m_{k_2}\}$ is the normalized fraction of signals ($n_{com}^{k_1,k_2}$) in common between the two groups. The measure takes the form of a reversed sigmoid function on the compact support $\beta_{com}^{k_1,k_2} \in [0,1]$, as shown in Figure 2. It is such that, high pairwise *input* diversity values are assigned to those pairs of groups whose fraction of common signals is relatively low (e.g. lower than 30%), whereas it penalizes group pairs with too many signals in common (e.g. more than 50%).

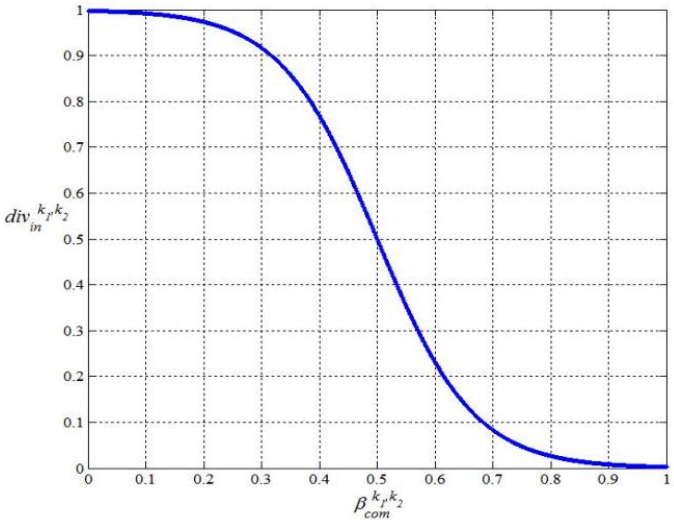

Figure 2. Group pairwise *input* diversity function, Eq. (1)

In the special case that two groups have the same number of signals, i.e., $m_{k_1} \equiv m_{k_2} = m$, then $\beta_{com}^{k_1,k_2}$ is 1 (and thus $div_{in}^{k_1,k_2} = \frac{1}{1+e^6} \approx 0$) if the two groups contain the exact same signals and $\beta_{com}^{k_1,k_2}$ is 0 (and thus $div_{in}^{k_1,k_2} = \frac{1}{1+e^{-6}} \approx 1$) if the two sets of signals are completely disjoint; if, instead, the sizes of the two groups differ, then the maximum number of signals in common for the two groups is $n_{com,MAX}^{k_1,k_2} = \min\{m_{k_1}, m_{k_2}\}$ and thus $\beta_{com,MAX}^{k_1,k_2} < 1$. This well represents the fact that even if a group k_1 is completely included in another one k_2, i.e. $k_1 \subset k_2$, their pairwise *input* diversity is not zero due to the presence in k_2 of some signals not included in the smaller set k_1.

To compute *input* diversity at the level of the ensemble of groups, first the diversity for each signal $i = 1,2,...,n$ is calculated. Considering the generic signal i included in K_i groups, the signal *input* diversity d_{in}^i is taken as the average of its K_i groups' pairwise diversities $div_{in}^{k_1,k_2}$, $k_1, k_2 = 1,2,...,K_i$, $k_1 \neq k_2$, viz.:

$$d_{in}^i = \frac{1}{K_i} \sum_{k_1=1}^{K_i} \left(\frac{1}{K_i - 1} \sum_{\substack{k_2=1 \\ k_2 \neq k_1}}^{K_i} div_{in}^{k_1,k_2} \right) \tag{2}$$

The ensemble input diversity δ_{in} is, then, simply computed as the average of the signals diversities:

$$\delta_{in} = \frac{1}{n} \sum_{i=1}^{n} d_{in}^i \tag{3}$$

3.2 Injecting diversity in the training data sets through BAGGING

Further diversity can be injected in the ensemble of models by training them on different data sets. In this work, this is achieved by the BAGGING technique which has proved successful in many applications [23, 27].

First of all, the data set $\mathbf{X}$ of N patterns available is partitioned into a training set $\mathbf{X}_{TRN}$ (made of N_{TRN} patterns) and a test set $\mathbf{X}_{TST}$ (made of N_{TST} patterns). The former is used to train the individual models, whereas the latter is used to verify the ensemble performance in the reconstruction task.

In general, BAGGING amounts to generating a number N_B of bootstrapped replicates $\mathbf{X}_{trn}^{B,h}$, $h = 1,2,...,N_B$ of the training set $\mathbf{X}_{trn}$ by randomly sampling (with replacement) for each replicate a fraction $\theta_B \in (0,1]$ of the total number of training patterns N_{trn}. If the fraction θ_B is large, the individual BAGGING training sets overlap significantly and the probability of not including a training pattern in any of the BAGGING training sets is very small, so that all the training patterns are likely to appear in at least one BAGGING training set and some patterns will appear multiple times in a given set; instead, if the fraction is small, some BAGGING training sets can be completely disjoint and some training patterns might not appear in any BAGGING training set.

3.3 Verifying (output) diversity in the ensemble of models

Figure 3 illustrates the combination of the RFSE and BAGGING techniques to inject diversity in the ensemble. As a result of the RFSE, K signal-diverse groups are identified, upon which K PCA models are constructed; BAGGING then proceeds to sampling (with replacement) a number $N_B = K$ of different training sets $\mathbf{X}_{trn}^{B,k}$, $k = 1,2,...,K$, each constituted by a fraction θ_B (equal for all the BAGGING subsets) of the original number of training patterns N_{trn}, i.e. $N_{trn}^{B} = \theta_B N_{trn}$; finally, the generic k-th model based on the signals of group k, randomly selected by RFSE, is trained with the set of data $\mathbf{X}_{trn}^{B,k}$, randomly sampled by BAGGING, for $k = 1,2,...,K$.

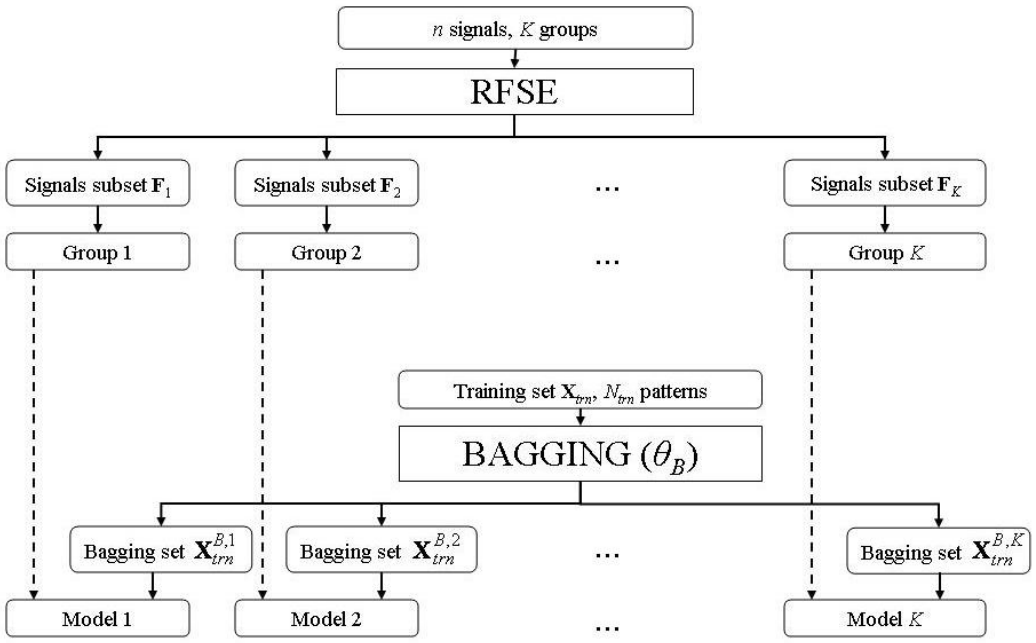

Figure 3. Combined scheme of the RFSE and BAGGING techniques to inject diversity in the ensemble

The total amount of diversity injected in the ensemble, hereby called *output* diversity, is the result of the combination of the RFSE and BAGGING randomizations.

To verify the amount of *output* diversity effectively injected in the ensemble, a measure is here proposed, which is directly related to the signal reconstruction performances obtained by the models of the ensemble. Let us consider a generic signal i included in K_i groups, $i=1,2,...,n$. For a given test pattern t, different from those used for the training of the K_i models, each group k provides the individual reconstruction $\hat{f}_i^k(t)$ of the signal $f_i(t)$, $k=1,2,...,K_i$, $t=1,2,...,N_{tst}$. For an accurate and robust reconstruction of the test values, the groups' reconstructions, or, analogously, their associated reconstruction errors, must be diverse. In this sense, let $\varepsilon_i^k(t)=\hat{f}_i^k(t)-f_i(t)$, $k=1,2,...,K_i$, $t=1,2,...,N_{tst}$, be the reconstruction error on the t-th test pattern of signal i by group k. If the errors $\varepsilon_i^k(t), k=1,2,...,K_i$ have all the same magnitude and sign, then their aggregation (i.e., the ensemble reconstruction of the t-th point of signal i) will be affected by a bias and will not be accurate; on the other hand, if errors are diverse in magnitude and sign (i.e. if the groups provide diverse reconstructions), their combination provides a more accurate estimation of the test pattern of the signal.

In practice, this means requiring errors to have different signs and magnitudes distributed around zero, so that, from a mathematical point of view, this favourable situation for the generic signal i could be described in terms of a quasi-normal Gaussian distribution, with mean equal to zero and standard deviation σ, of the K_i groups' reconstruction errors of each t-th test value, i.e. $\left\{\varepsilon_i^{k_1}(t),\varepsilon_i^{k_2}(t),...,\varepsilon_i^{K_i}(t)\right\}\sim N(0,\sigma)$.

On this basis, we define a procedure to measure the *output* diversity of the K_i groups in reconstructing the generic point t of signal i. First, the empirical cumulative distribution function (cdf) $F_{K_i}^t(\varepsilon)$ of the groups' reconstruction errors $\varepsilon_i^k(t)$, $k=1,2,...,K_i$, is computed:

$$F_{K_i}^t(\varepsilon)=\frac{1}{K_i}\sum_{k=1}^{K_i} I_{\varepsilon_i^k(t)\le\varepsilon} \tag{4}$$

where $I_{\varepsilon_i^k(t)\le\varepsilon}$ is the indicator function equal to 1 if $\varepsilon_i^k(t)\le\varepsilon$ and 0 viceversa.

The empirical (real) cdf of the reconstruction errors is compared with the (ideal) cdf of a quasi-normal Gaussian distribution $F_G(\varepsilon)$ of mean zero and given σ and the maximum absolute distance $D_t=\max_{k=1,2,...,K_i}\left|F_{K_i}^t(\varepsilon_i^k(t))-F_G(\varepsilon_i^k(t))\right|$, $D_t\in[0,1]$, between the two cdf's is computed (Figure 4) [2]. A small distance D_t corresponds to a good distribution of the reconstruction errors, whereas a large distance reveals the presence of a bias in the reconstruction of the t-th signal's value.

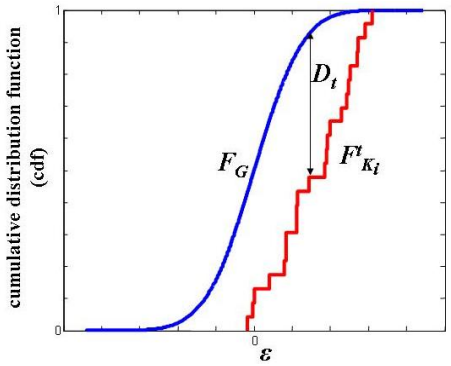

Figure 4. Comparison in terms of maximum absolute distance (D_t) between the empirical $(F_{K_i}^t)$ and Gaussian (F_G) cdfs.

[2] Operatively and without loss of generality, we consider an ideal Normal distribution $F_G(\varepsilon)=N(0,1)$ and, for an honest comparison, we pre-normalize the group reconstruction errors $\varepsilon_i^k(t), k=1,2,...,K_i$ in order to have a standard deviation equal to 1 in the real distribution $F_{K_i}^t(\varepsilon)$.

The pointwise *output* diversity between the K_i groups in reconstructing the t-th value of the i-th signal is simply computed as $div_i^t = 1 - D_t$. The *output* diversity d_{out}^i for signal i is taken as the average of the pointwise diversities:

$$d_{out}^i = \frac{1}{N_{tst}} \sum_{t=1}^{N_{tst}} div_i^t \tag{5}$$

Finally, the ensemble *output* diversity is the average of the signals *output* diversities:

$$\delta_{out} = \frac{1}{n} \sum_{i=1}^{n} d_{out}^i \tag{6}$$

3.4 Determining the group size and number

In this work, a wrapper optimization of the ensemble performance is carried out in order to determine the group size m_{opt} (i.e. the optimal number of signals to have in each group, equal for all groups in the ensemble) and the ensemble size K_{opt} (i.e. the number of groups in the ensemble, each one constituted by m_{opt} signals).

The optimal group size m_{opt} is a problem-dependent parameter [27], which is chosen within a pre-defined range of group sizes, i.e. $m_{opt} \in [m_{\min}, m_{\max}]$. A crude, exhaustive wrapper approach is here adopted. Operatively, for each candidate group size $m \in [m_{\min}, m_{\max}]$, an ensemble of K groups of m signals is generated by randomly sampling the signals of the groups by RFSE, K corresponding PCA-based reconstruction models are trained on different data sets obtained by BAGGING and the ensemble reconstruction performance is computed. Since the scope of the work is to provide a robust ensemble of groups, the ensemble performance is computed also on a set of test signals artificially disturbed (see Section 4.1 for details on the disturbance procedure). The optimal group size m_{opt} is the one corresponding to the ensemble of groups providing the best reconstruction performances on both undisturbed and disturbed signals. Notice that searching for an optimal group size equal for all groups is necessary in practice to reduce the number of parameters to be optimized.

A relevant issue for robust signal reconstruction is that each signal i be included in a suitable number of groups $K_i >> 1$, i.e. to have appropriate redundancy of signals representation in the ensemble. The average signal redundancy in the groups of the ensemble can be computed as:

$$R = \frac{1}{n} \sum_{i=1}^{n} K_i \tag{7}$$

Considering that different overlapping groups have some signals in common, or, dually, different signals have some groups in common, the total number of signals (with repetitions) in the ensemble of groups can be expressed either by summing the number of signals m_k constituting each group $k = 1,2,...,K$ (i.e., the groups' sizes) or by summing the number of groups K_i including each signal $i = 1,2,...,n$ (i.e., the signals' redundancies), viz.:

$$\sum_{k=1}^{K} m_k = \sum_{i=1}^{n} K_i \tag{8}$$

In the case in which each group includes the same number of signals (i.e., $m_k \equiv m$, $\forall k$) and each signal is included in the same number of groups (i.e., $K_i \equiv R$, $\forall i$), Eq. (8) becomes:

$$K \cdot m = n \cdot R \tag{9}$$

For a given (optimal) group size m_{opt}, the (optimal) number of groups K_{opt} in the ensemble can be determined by fixing the desired (optimal) signal redundancy R_{opt}, viz.:

$$K_{opt} = \frac{n}{m_{opt}} R_{opt} \tag{10}$$

The K_{opt} groups, each one constituted by m_{opt} signals, are then generated resorting to the RFSE technique and the corresponding PCA-based reconstruction models are BAGGING-trained.

Figure 5 sketches the scheme of the overall algorithm for devising and verifying the ensemble of models for signal reconstruction.

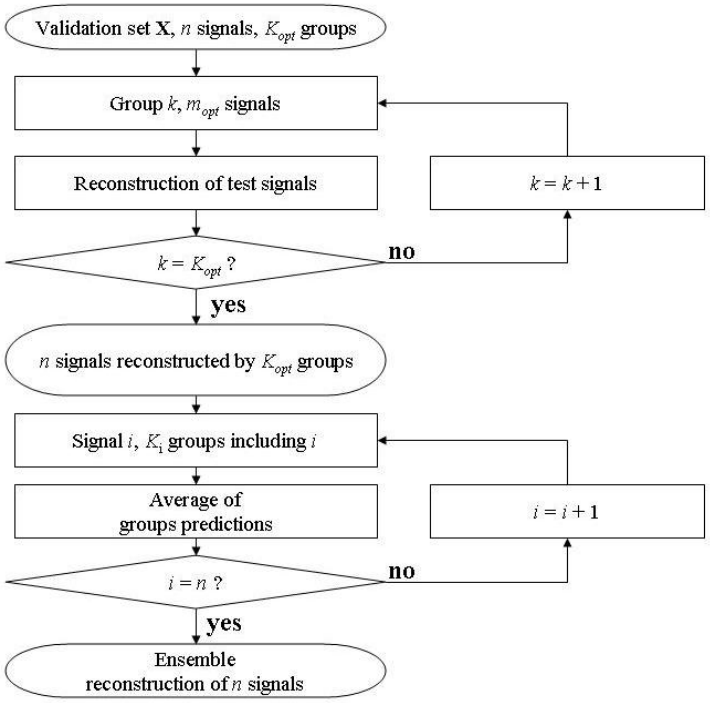

Figure 5. Sketch of the ensemble algorithm for signal reconstruction

3.5 Combination of the outcomes of the models in the ensemble

The combination of the outcomes of the ensemble of models is performed by simple averaging [13, 22, 23]. This way of combining the models outputs can be seen as an extension to a regression problem of the Simple Voting (SV) technique adopted in classification problems to combine the class assignments of the single classifiers constituting the ensemble [14, 28].

When the N_{TST} patterns of the test set are fed in input to the generic k-th PCA model, based on the m_{opt} signals of group k and trained on the N_{TRN}^{B} BAGGED patterns, this gives in output the predictions $\hat{f}_i^k(t)$, $t = 1,2,...,N_{TST}$, $i = 1,2,...,m_{opt}$. The ensemble reconstruction of the N_{TST} patterns of the generic i-th signal, $\hat{f}_i^E(t)$, $t = 1,2,...,N_{TST}$, $i = 1,2,...,n$, is then obtained by averaging of the predictions $\hat{f}_i^k(t)$ of the K_i groups including signal i:

$$\hat{f}_i^E(t) = \frac{1}{K_i} \sum_{k=1}^{K_i} \hat{f}_i^k(t) \tag{11}$$

To evaluate the ensemble performance, first the absolute signal reconstruction error is computed as[3]:

[3] In the application that follows, each signal of the validation set has been previously normalized in the range [0.2, 1], for convenience.

$$\varepsilon_i^E = \frac{1}{N_{TST}} \sum_{t=1}^{N_{TST}} \left| f_i(t) - \hat{f}_i^E(t) \right| \tag{12}$$

Then, an ensemble performance index is computed as the average of the absolute signal reconstruction errors:

$$\eta^E = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^E \tag{13}$$

4. Application

The ensemble approach has been applied to a real case study concerning the validation and reconstruction of 215 signals measured at a nuclear Pressurized Water Reactor (PWR) located in Loviisa, Finland. A total number N =12713 of 215-dimensional patterns is available, $f_1(t), f_2(t),..., f_i(t),..., f_{215}(t)$, $t = 1,2,...,N$. Data signals have been sampled every hour from February 28, 2006 to November 1, 2007 from a corresponding number of sensors and include the measurements related to two plant outages occurred in the periods from June 26 to October 1, 2006 and from September 8 to September 23, 2007, respectively. A set $\mathbf{X}_m$ constituted by $N_m = 6000$ patterns is used for the *wrapper* determination of the groups' optimal size m_{opt}; a set $\mathbf{X}_{RB}$ constituted by the remaining $N_{RB} = 6713$ patterns is used for the *wrapper* determination of the ensemble size K_{opt} and the BAGGING fraction θ_B. Notice that the patterns of the transient related to the first outage are included in $\mathbf{X}_m$, while those of the transient of the second outage are in $\mathbf{X}_{RB}$. The PCA models (see Appendix) have been constructed with the code http://lib.stat.cmu.edu/multi/pca, adapted to perform the signal reconstruction task of interest here.

4.1 Determination of m_{opt}

The optimal size m_{opt} of the ensemble's groups is sought in the range of values $m = [20,40]$. The group dimensions in this range are expected to lead to accurate and robust regression models, while being at the same time easy to handle from a computational viewpoint. For verification purposes, the search has been further stretched up to group sizes $m_{\max} = 70$.

Operatively, for each candidate group size $m = 20, 21, ...$, an ensemble constituted by $K = 150$ groups is generated by randomly sampling m signals in each group. The ensemble performance is then computed following the scheme illustrated in Figure 5 and the ensemble error η^E is computed as Eq. (13).

A robust ensemble of models is expected to be able to reconstruct the signals from faulty sensors, e.g. due to random noises, offsets or drifts. In other words, when a faulty signal is sent in input to the PCA models which include that signal, their ensemble should still provide in output a good estimate of the true value of the signal, by exploiting the information coming from the non-faulty signals in the groups of the ensemble.

For verification purposes, a partially faulty test set has been generated by introducing anomalies in a fraction of the patterns. More precisely, the signal values of a generic test pattern are altered by a random noise (with probability $p^{RN} = 0.02$) or by setting them equal to the offset value of the corresponding sensor (with probability $p^{OF} = 0.01$); with probability 0.97 they are not affected at all. The ensemble of groups is then tested on this test set of disturbed signals and the corresponding performance η^{E*} is computed as Eq. (13).

Actually, to account for the variability of the training and test sets, five ensembles (each constituted by $K = 150$ groups with randomly sampled signals) have been generated for each group size m and for each of the five ensembles the computation of the accuracy and robustness indicators has been 5 times cross-validated. The overall ensemble performance associated to the generic group size m is finally computed as the average over all cross-validations of all different ensembles.

Figure 6 shows the results of the analysis. In general, ensembles built on large groups are more accurate on undisturbed signals (smaller η^E values, Figure 6 left), whereas those built using both small $(m < 30)$ and large $(m > 60)$ groups have slightly worse performances on disturbed signals (larger η^{E*} values, Figure 6 right), i.e. they are less robust. Notice that,

despite the performance differences are very small, at least for the case of undisturbed signals, there is a trend revealing a tentative optimal group size.

Small values of m_{opt} have not been considered since in general reconstruction models built with few signals are not robust, i.e. if one or more signals are disturbed or missing the reconstruction of all the others is negatively affected. For this reason, within the range [20-40], $m_{opt} = 38$ shows the best compromise between accuracy and robustness and is chosen as the ensemble groups' optimal size.

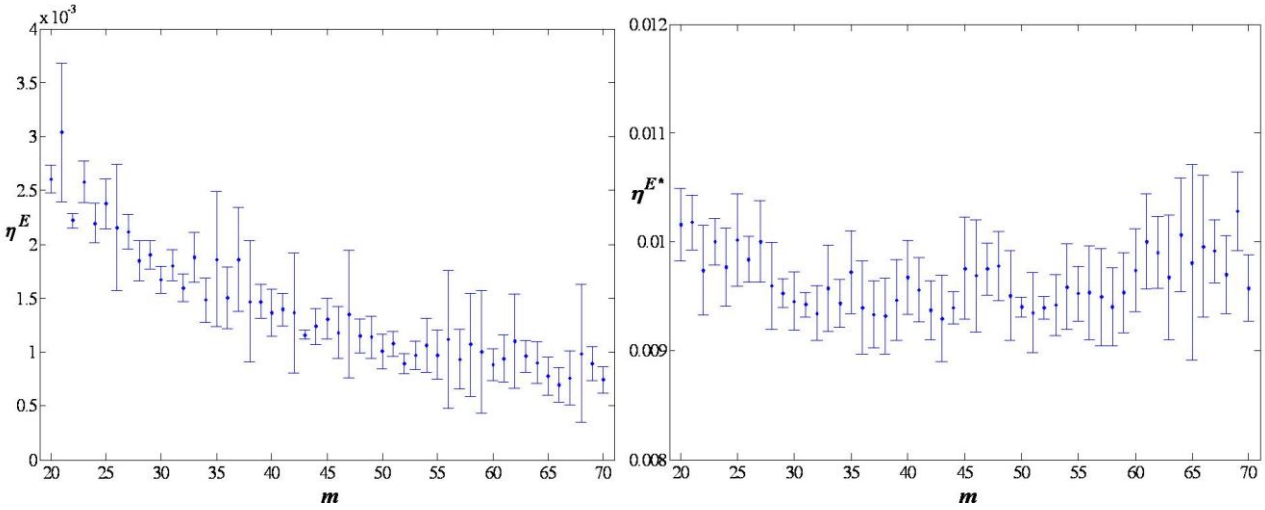

Figure 6. Performances of the ensemble for different group sizes (m) on undisturbed signals (η^E , left) and disturbed signals (η^{E*} , right). Cross-validation statistical errors (one standard deviation above and below the average) are also reported in the Figure.

4.2 Determination of K_{opt} and θ_B

Once the optimal group size m_{opt} has been fixed, it is possible to proceed to determine the optimal ensemble size K_{opt} and BAGGING fraction θ_B. As explained in Section 3.4, the ensemble size is obtained by the empirical relation of Eq. (9), for a given desired signal redundancy R_{opt}. Eight candidate values for R_{opt} have been chosen, based on an expert engineering evaluation; correspondingly, eight candidate ensemble sizes K_{opt} have been identified. These values are reported in Table 1, together with the actual mean signal redundancy (computed as Eq. 7) and standard deviation obtained after the groups have been created with the RFSE procedure. For the BAGGING fractions, seven values have been chosen, namely 1, 0.8, 0.6, 0.4, 0.1, 0.02 and 0.005; the case of no BAGGING procedure (in which all the K_{opt} models are trained using the same data set) has also been considered.

Required redundancy, R_{opt}	Ensemble size, K_{opt}	Actual redundancy, R
10	57	10.07 ± 3.12
15	85	15.02 ± 3.91
20	114	20.14 ± 4.67
25	142	25.09 ± 5.30
30	170	30.04 ± 6.05
40	227	40.12 ± 6.98
50	283	50.01 ± 7.54
60	340	60.09 ± 8.99

Table 1. Pre-defined, desired values for R_{opt}, corresponding ensemble sizes K_{opt} and actual redundancies R.

The results of the joint optimization of K_{opt} and θ_B are reported in Figure 7 in terms of ensemble accuracy, i.e. reconstruction errors on undisturbed signals (top graph), and robustness, i.e. reconstruction errors on disturbed signals (bottom graph).

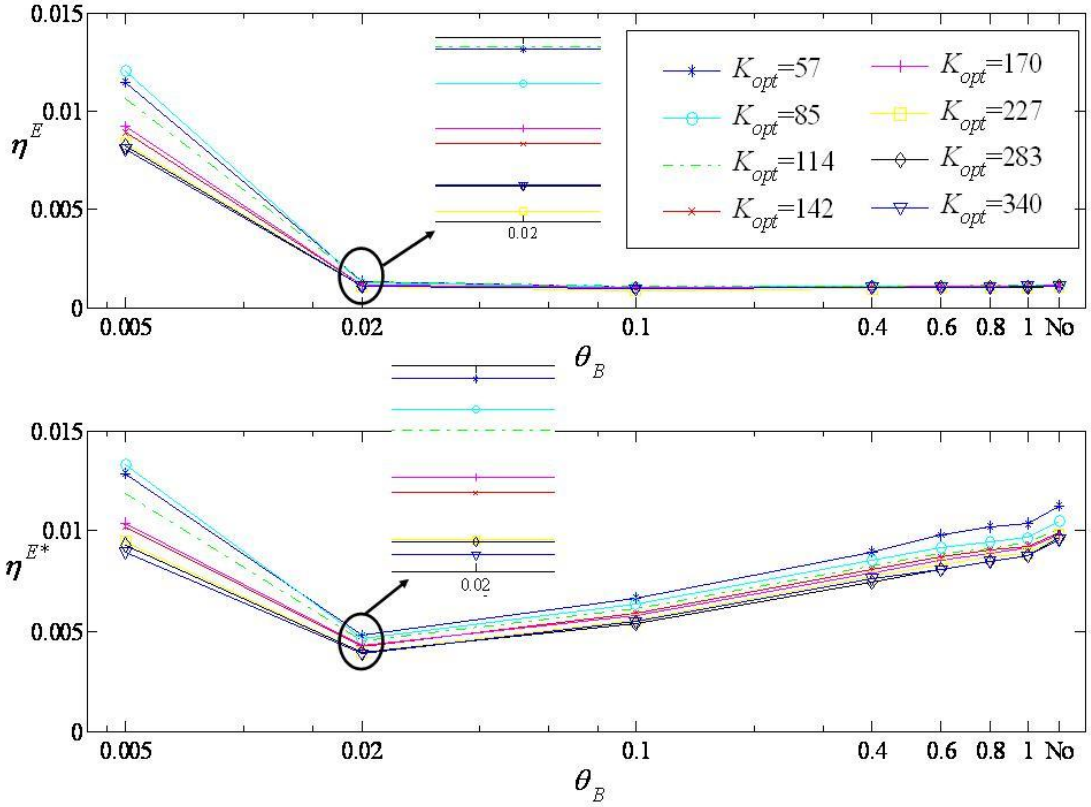

Figure 7. Results of the joint optimization for determining the optimal ensemble size K_{opt} and bagging fraction θ_B in case of undisturbed (top graph) and disturbed (bottom graph) signals.

Concerning the ensemble size, in general, high signal redundancy and, correspondingly, large numbers of groups guarantee better accuracy and robustness. Nevertheless, using a large number of groups in the ensemble leads to a considerable increase in the computational cost of training and testing the models, with only a slight improvement of the ensemble performances.

Indeed, the error scales are small (especially for the undisturbed signals,) and the error bars (which have not been inserted for visual clarity) are superposed. Nevertheless, from an operative point of view, a decision on how many groups to use and which bagging fraction to adopt must be made. For this reason, $R_{opt} = 25$ is the value chosen for the desired signal redundancy and, correspondingly, $K_{opt} = 142$ is the fixed ensemble size. This choice allows having good ensemble performances with a low computational cost and furthermore it allows having an honest comparison with the genetic algorithm-based optimized grouping [18] which is presented in Section 4.4.

Concerning the BAGGING procedure, in general the results show that injecting diversity by training the models on different data sets improves the ensemble performances. Furthermore, the reconstruction errors decrease (especially for disturbed signals) if the fraction of sampled training patterns N_{trn}^{B} in the bagging training sets becomes smaller. In fact, by so doing, many BAGGING training sets are completely disjoint and therefore highly diverse.

The minimum reconstruction error is obtained for $\theta_B = 0.1$ in the case of undisturbed signals and $\theta_B = 0.02$ in the case of disturbed data. The remarkable decrease of the ensemble error on the disturbed signals when choosing $\theta_B = 0.02$ instead of $\theta_B = 0.1$ leads us to retain $\theta_B = 0.02$ as the optimal fraction, despite a slight worsening of the ensemble performances

on the undisturbed set. Such a low value can be explained by the fact that excluding many training patterns from the signal data set allows removing with high probability those spurious signal measurements (e.g. spikes) that usually compromise the reconstruction of all the signals in the single groups.

Finally, if the fraction becomes too small (e.g., $\theta_B = 0.005$), the ensemble performances considerably worsen due to the lack of data representation in the BAGGING training sets.

4.3 BAGGING

This Section intends to illustrate specific situations in which Bagging is and is not effective in terms of the output diversity. Indeed, the BAGGING procedure is intended to enhance the *output* diversity of the ensemble. According to the adopted *output* diversity measure for the generic signal i, the models built using the signals of its K_i groups are diverse if they make different errors in reconstructing the signal. More precisely, it is conjectured that the most beneficial diversity is achieved if the reconstruction errors for each signal's test point are distributed as a quasi-normal Gaussian function (Section 3.3).

In order to show the effects of BAGGING, two test patterns of signal 19 included in $K_{19} = 25$ groups are here analyzed in details: the first (#196) is a signal measurement taken during normal plant operation, whereas the other (#731) corresponds to a measurement during plant shut-down. In Figure 8, the effects of BAGGING on the errors distribution of the 25 groups including signal 19 on the two test patterns are reported in both cases of undisturbed (top graphs of Figure 8) and disturbed (bottom graphs of Figure 8) patterns. When the pattern is disturbed, performing BAGGING allows obtaining more diverse groups predictions (i.e. the groups' errors distribution is more similar to that of a quasi-normal Gaussian) and thus a more accurate and robust reconstruction of the pattern's correct value. On the other hand, if the pattern is undisturbed, performing BAGGING can degrade the ensemble performance by reducing the groups diversity in the reconstruction of that pattern, i.e. by negatively affecting the distribution of the groups' errors and rendering it quite dissimilar to the Gaussian cumulative distribution, as for pattern 196 (top-left graph).

Furthermore, looking at the results for pattern 731 (corresponding to the shut-down transient) one sees that the ensemble trained without BAGGING does not become aware of the disturbance affecting it (bottom-right graph) and the errors distribution remains almost unaltered, as evident from the comparison between the top-right and bottom-right graphs; instead, BAGGING allows recognizing the occurrence of the disturbance and the distribution of the models' errors is adjusted to become more similar to a quasi-Gaussian distribution. In this view, the superior performance in the reconstruction of pattern 731 without disturbance obtained with BAGGING (top-right) can be explained by considering the transient part of the signal as disturbances, on whose reconstruction BAGGING has certainly positive effects.

In this sense, bagging has proved effective in improving the performances on disturbed data more than on the undisturbed ones.

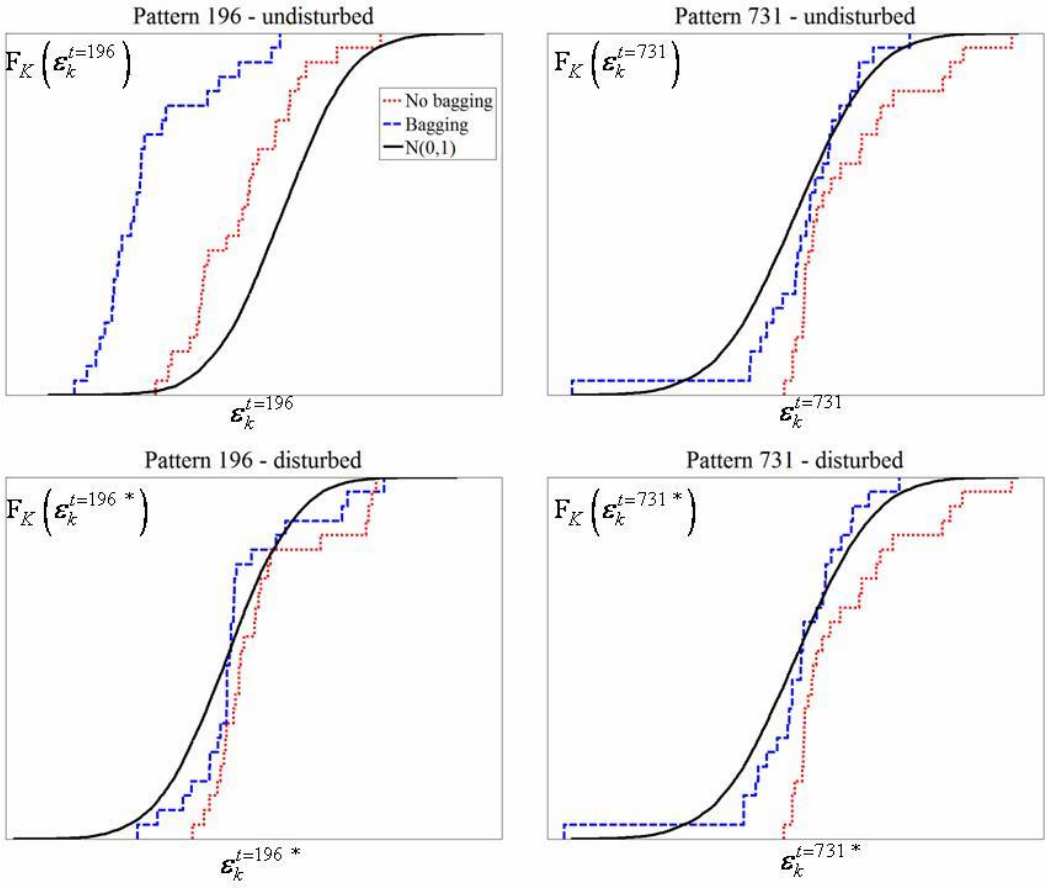

Figure 8. Effects of BAGGING on the distribution of groups errors in reconstructing test pattern number 196 (left graphs) and 731 (right graphs), when undisturbed (top graphs) and disturbed (bottom graphs).

4.4 Comparison with the MOGA approach to signal grouping

In order to evaluate the advantages and limitations of the proposed ensemble approach, a comparison to the signal grouping based on Multi-Objective Genetic Algorithm (MOGA) optimization investigated in [18] is proposed in this Section. The comparison is made on an ensemble of 150 groups.

In the MOGA approach, the search for the set of $K_{MOGA} = 150$ optimal groups of signals has been iteratively carried on by optimizing (i.e. maximizing) the correlation between the groups' signals and the *input* diversity between the groups (the interested reader may refer to [18] for a fully detailed explanation). The comparable ensembles size allows for an honest comparison of the two techniques.

Concerning the *input* diversity δ_{in} (Eq. 3), i.e. the average degree of group overlapping in terms of signals, the RFSE shows a considerable improvement with respect to the MOGA approach ($\delta_{in}^{RFSE} = 0.8884$, whereas $\delta_{in}^{MOGA} = 0.6750$). This proves that, according to the adopted measure, randomly selecting the signals in the groups corresponds to a high injected diversity. As illustrated in Figure 2, the RFSE groups (all with $m_{opt} = 38$ signals) have on average 31% of signals in common, whereas the MOGA groups (ranging from 8 to 147 signals), show on average 43% of signals in common due to the presence of small groups which tend to be more easily included in large groups.

Figures 9 and 10 report the ensemble reconstruction errors (Eq. 13) and *output* diversities (Eq. 6), respectively, computed for undisturbed (η^{E}, δ_{out}) and disturbed (η^{E*}, δ_{out}^{*}) signals, using the PCA models based on the $K_{opt} = 142$ and the $K_{MOGA} = 150$ groups obtained from the wrapper randomized (by RFSE) and MOGA techniques, respectively. The two ensembles are compared with and without performing BAGGING.

Enhancing *input* diversity with the RFSE approach improves the robustness of the ensemble on disturbed signals (η^{E*}, Figure 9) and corresponds to an increase of the *output* diversity on disturbed signals (δ^{*}_{out}, Figure 10). Nevertheless, the *output* diversity on undisturbed signals (δ_{out}, Figure 10) of the RFSE groups without BAGGING is lower and correspondingly the performance is worse (η^{E}, Figure 9). This performance degradation can be explained by the fact that, while in the MOGA the signals in the groups are selected based on their mutual correlation (a characteristic conjectured to be related to their capability of regressing one another [11, 12, 16-18]), the signals in the RFSE groups are randomly selected regardless of their mutual correlation or any other criterion for optimizing their reconstruction capabilities.

The robustness of both RFSE and MOGA ensembles (η^{E*}, Figure 9) is considerably improved when BAGGING is applied as indicated by the increase of the corresponding ensemble *output* diversities on disturbed signals (δ^{*}_{out}, Figure 10). Nevertheless, when performing BAGGING on the RFSE ensemble, the increase of the *output* diversity on undisturbed signals (δ_{out}, Figure 10) is not followed by an increase on the reconstruction performances on undisturbed data (η^{E}, Figure 9). In general, performing BAGGING degrades the model accuracy in reconstructing undisturbed signals.

Finally, notice that BAGGING slightly contributes to RFSE at increasing the undisturbed error η^{E}, whereas the effect on η^{E} due to RFSE is more evident when comparing RFSE with MOGA. This means that too much diversity randomly introduced in the groups does not improve the capability of reconstructing undisturbed signals, which instead comes from having highly correlated signals (as in the MOGA groups); on the other hand, a robust signal reconstruction in case of disturbances is due to high models diversity (as in the RFSE and MOGA approaches with BAGGING).

The robustness of the RFSE and MOGA ensembles has been then specifically tested for comparison on the reconstruction of faulty signals in case of multiple sensor failures. Ten signals have been chosen as objects of the analysis. Approximately, the first half of the signals has been left undisturbed as in the normal operation, while, in order to simulate a sensor failure, a linear drift decreasing the values of the signals up to 25% of their real values has been introduced in the remaining test values. The validation set has been linearly divided into training and test only once, i.e. without cross-validation.

Figure 11 shows the results of the reconstruction of signal 214 (electrical power) obtained by the RFSE and the MOGA approaches both trained with BAGGING procedure ($\theta_B = 0.02$). When the signal is undisturbed the highly correlated signals of the (less diverse) MOGA groups allow for a more accurate reconstruction, whereas as soon as the drift begins the more diverse RFSE groups are capable of providing a more robust signal reconstruction.

Finally, in order to give an overview of the advantages and limitations of the two approaches, Table 2 summarizes the characteristics of the signal grouping structure and the ensemble reconstruction performances obtained by the RFSE and MOGA approaches, respectively.

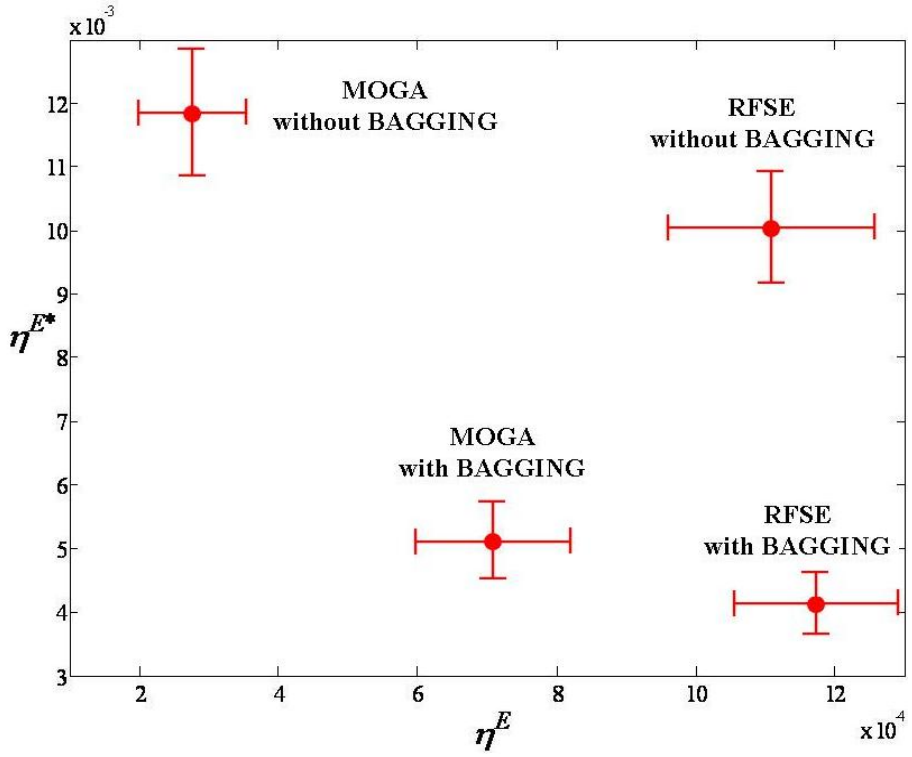

Figure 9. Comparison of the MOGA and RFSE approaches to signal grouping in terms of the ensemble reconstruction errors on undisturbed (η^{E}) and disturbed (η^{E*}) signals, with and without performing BAGGING

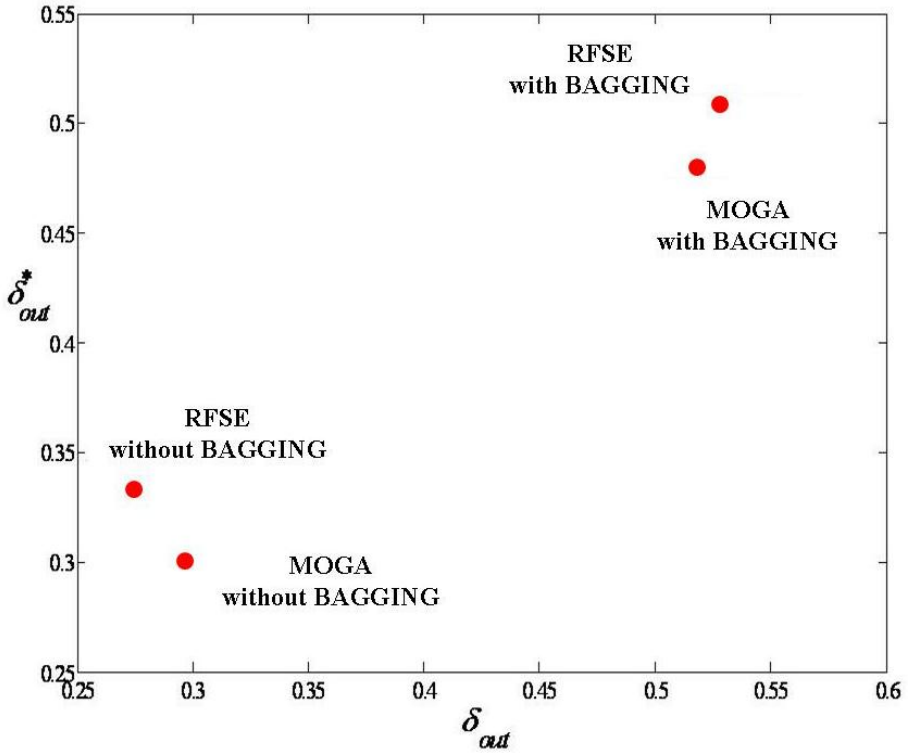

Figure 10. Comparison of the MOGA and RFSE approaches to signal grouping in terms of the ensemble *output* diversity measured when regressing undisturbed (δ_{out}) and disturbed (δ^{*}_{out}) signals, with and without performing BAGGING

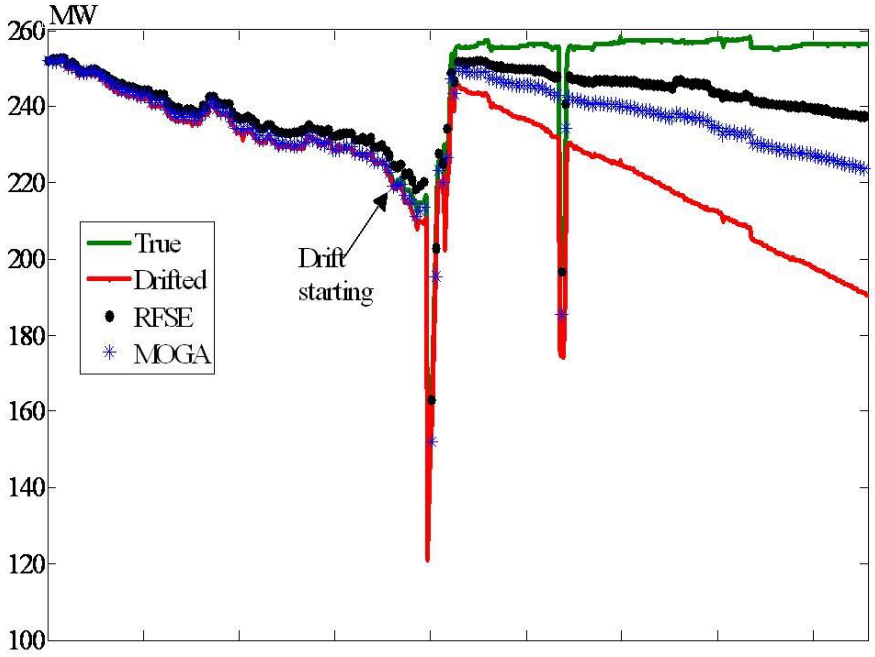

Figure 11. RFSE (dots) and MOGA (stars) ensemble reconstruction of signal 214 (light line) when partly affected by a linear drift (dark line)

		RFSE	MOGA[11]
Signal grouping	*Optimized parameters*	Group size, signal redundancy and number of groups (wrapper approach).	Intra-group signals correlation and inter-group signals diversity (filter approach).
	Group size, m	Controlled by wrapper optimization accounting for accuracy and robustness. The same (38) for all groups.	Neither controlled, nor optimized. Groups of different sizes ranging from 8 to 147 signals.
	Signal redundancy, R	Controlled *a priori* and optimized based on accuracy and robustness. Maintained on average after selecting randomly the groups' signals and evenly distributed among the signals.	Not imposed *a priori*, but indirectly optimized by maximizing signals diversity in the groups. Unevenly distributed among the signals.
	Number of groups (ensemble size), K	Set as a function of the optimized group size and signal redundancy.	Set *a priori* according to the dimensions of the solution search space. Strongly influencing the computational cost of the optimization.
	Computational cost	Medium-low; dependent on the type of regression model adopted to run the wrapper optimization of the ensemble dimension parameters, m_{opt} and K_{opt}.	High; strongly related to the ensemble dimension parameters, i.e. the number of signals and groups involved in the search.
Ensemble signal validation	*Ensemble input diversity*	Very high; due to signals random selection.	Medium-high; optimized by enhancing diversity between the groups, but indirectly reduced by maximizing signals correlation.
	Accuracy (reconstruction of undisturbed signals)	Low; due to lack of optimization of groups individual properties (e.g. mutual correlation).	High; due to high signals correlation in the groups achieved during the optimization.
	Ensemble output diversity	High; mostly due to model diversity injected with the bagging technique, but also increased by the combined injection of signals diversity (with random selection).	Medium; due only to the BAGGING procedure.
	Robustness (reconstruction of disturbed signals)	High; due to the presence of diverse information both in the groups' signals (RFSE) and in the training data sets (BAGGING).	Medium; due only to the BAGGING procedure.

Table 2. Summary of the characteristics and performances of the RFSE and MOGA approaches

4.5 Validation of the approach on two additional case studies

The overall RFSE approach has been verified on two different case studies for validation. The first concerns 84 signals measured at a Swedish nuclear Boiling Water Reactor (BWR) situated in Oskarshamn, the other the reconstruction of 920 simulated signals of the Swedish Forsmark-3 Boiling Water Reactor (BWR). The data available for the measured 84 signals have been sampled every 10 minutes from May 31, 2005 to January 5, 2006 from a corresponding number of sensors, providing a total amount of $N = 30080$ time samplings. Regarding the 920 signals, the sensors measurements

have been simulated with a sampling rate of one hour, under conditions of reactor start up, normal operation and shut down.

Table 3 briefly reports the characteristics of the data sets and the grouping parameters set with the same wrapper optimization hereby proposed. Notice that in order to limit the number of groups in the ensemble which greatly affects the computational cost R_{opt} for the Forsmark-3 case study has been set very low. Notice also that the value of the optimal θ_B is higher for these case studies. This is due to the fact that data contain a smaller number of spurious measurements (especially the simulated data). The results of the ensemble signal reconstruction in terms of disturbed and undisturbed errors with and without BAGGING are also reported in Table 3; the performances resemble those achieved in the previous case study.

		Oskarshamn case study	Forsmark-3 case study
Number and type of signals		84 measure signals	920 simulated signals
Number of available measurements, N		30080	5463
Number of signals per group, m_{opt}		25	70
Required redundancy, R_{opt}		15	7
Number of groups in the ensemble, K_{opt}		51	92
Optimal bagging fraction, θ_B		0.25	0.4
Undisturbed reconstruction error, η^E	No BAGGING	8.23×10^{-4}	5.53×10^{-4}
	BAGGING	9.44×10^{-4}	7.71×10^{-4}
Disturbed reconstruction error, η^{E*}	No BAGGING	7.11×10^{-3}	3.39×10^{-3}
	BAGGING	3.34×10^{-3}	1.28×10^{-3}

Table 3. Characteristics, parameters and ensemble performances of the Oskarshamn and Forsmark-3 case studies.

Finally, the results of the Oskarshamn case study without Bagging can be compared with those achieved by using the Multi-Objective Genetic Algorithm approach to signal grouping presented in [17]. The MOGA has been aimed at searching for 100 optimal groups of signals by maximizing the correlation between the groups' signals and the *input* diversity between the groups. The reconstruction errors on undisturbed (η^E) and disturbed (η^{E*}) signals obtained by the MOGA approach are 9.39×10^{-4} and 16.7×10^{-3}, respectively. Even though the comparison has been carried on a non-comparable ensemble size (51 and 100 groups for the RFSE and MOGA, respectively), still the results confirm the higher robustness of the RFSE approach in reconstructing disturbed signals.

In general, even though the methodology proposed in this work is obviously problem-dependent with respect to the optimal grouping parameters which must be discerned for the specific data set, these results show that it is applicable to any problem involving a large number of signals to be validated and reconstructed for which training a single model entails convergence problems (such as for the iterative training of Auto-Associative Neural Networks) and shows significantly less robustness.

5. Conclusions

In this work, a novel approach to the reconstruction of the signal values from faulty sensors has been proposed, based on an ensemble of PCA models. The set of sensor signals, too large to be handled effectively with one single reconstruction

model, is subdivided into small, overlapping groups and a PCA-based reconstruction model is developed for each group. The outcomes of the models are then combined by simple averaging to obtain the ensemble signal reconstruction.

The fundamental characteristic of diversity in the ensemble models has been generated by randomly sampling the signal features in the groups (by the RFSE procedure) and a fraction of the patterns for training the corresponding reconstruction models (by the BAGGING technique). The dimension parameters of the ensemble (i.e., the size of the groups and the number of groups in the ensemble) have been optimized by a direct wrapper approach.

The overall modelling scheme has been applied to the reconstruction of signals collected at a Finnish nuclear pressurized water reactor. Two additional case studies have been analyzed for validation purposes. The performances of the proposed approach have been compared with those of an equivalent ensemble obtained by means of a multi-objective genetic algorithm aimed at maximizing the intra-group signal correlation and the inter-group signal diversity. Considering the difficulties of calibrating the ensemble dimension parameters and of maintaining the group signals diversity in the MOGA optimization, and its not negligible computational cost, the RFSE approach has proved more efficient for large-scale applications involving hundreds of signals.

The approach proposed has shown considerable robustness in reconstructing signals when in presence of disturbances and drifts. At the same time, it allows to control and optimize the ensemble dimension (number of groups and of signals in the groups), which is fundamental for practical applications. On the other hand, not considering in any way the optimization of the models reconstruction capability is paid by a (small) loss of accuracy in reconstructing undisturbed signals. Finally, the performances of the overall RFSE approach have been verified on two different case studies providing similar satisfactory results and, in general, the methodology here presented has proved to be fast and robust and is currently object of further improvements.

The significance of the findings is to be framed within the practical problem tackled and goes beyond the numerical values of the results obtained. The ensemble approach proposed provides a feasible way for handling the validation and reconstruction of a set of a large number of signals,which cannot be handled by a single model. The method has proved its effectiveness, its low computational cost (an essential aspect for the application of any on-line sensor monitoring system) and its applicability to different signal sets. Finally, even though the numerical values show sometimes little improvements with respect to the reconstruction error, these must be seen in terms of enhanced robustness and plant production and safety when the validated and reconstructed signals are effectively used during operation. In this respect, notice the criticality to have a robust and accurate reconstruction of those signals that are used by the plant control systems, for which a small error in the reconstruction can lead to wrong control actions, possibly with significant production losses and safety threats: this last aspect is part of currently ongoing research.

References

[1] M. Hoffmann, “On-line Monitoring for Calibration Reduction”, HWR-784, OECD Halden Reactor Project, October 2005.

[2] M. Hoffmann, “Signal Grouping Algorithm for an Improved On-line Calibration Monitoring System”, Proceedings of FLINS, Genova, Italy, Aug. 2006.

[3] A. S. Heger, K. E. Holbert, A. M. Ishaque, Fuzzy Associative Memories for Instrument Fault Detection, 1996, Annals of Nuclear Energy, Vol. 23, No. 9, pp. 739-756.

[4] K. E. Holbert, A. S. Heger, A. M. Ishaque, Fuzzy Logic for Power Plant Signal Validation, 1995, Proceedings of the Ninth Power Plant Dynamics, Control & Testing Symposium, pp. 20.01-20.15, Knoxville, TN.

[5] X. Wang, K. E. Holbert, A Neural Network Realization of Linear Least-Square Estimate for Sensor Validation, 1995, Proceedings of the Ninth Power Plant Dynamics, Control & Testing Symposium, pp. 15.01-15.15, Knoxville, TN.

[6] K. E. Holbert, Neural Networks for Signal Validation in Nuclear Power Plants, 1992, Electric Power Research for the 90's, Proceedings of the Second Annual Industrial Partnership Program Conference.

[7] K. E. Holbert, B. R. Upadhyaya, An Integrated Signal Validation for Nuclear Power Plants, 1990, Nuclear Technology, Vol. 92, No. 3, pp. 411-427.

[8] P. F. Fantoni, M. I. Hoffmann, R. Shankar, E.L. Davis, On-line Monitoring of Instrument Channel Performance in Nuclear Power Plant using PEANO, 2003, Progress in Nuclear Energy, Vol. 43, No. 1-4, Pages 83-89.

[9] P. F. Fantoni , A. Mazzola, Multiple-Failure Signal Validation in Nuclear Power Plants using Artificial Neural Networks, 1996, Nuclear technology, Vol. 113, No. 3, pp. 368-374.

[10] D. Roverso, M. Hoffmann, E. Zio, P. Baraldi, G. Gola, "Solutions for plant-wide on-line calibration monitoring", 2007, Proc. ESREL 2007, Stavanger, Norway, Vol. 1, pp. 827-832.

[11] E. Zio, P. Baraldi, G. Gola, D. Roverso, M. Hoffmann, "Genetic Algoritms for Grouping of Signals for System Monitoring and Diagnostics", 2007, Proc. ESREL 2007, Stavanger, Norway, Vol. 1, pp. 833-840.

[12] P. Baraldi, E. Zio, G. Gola, D. Roverso, M. Hoffmann, "Genetic algorithms for signal grouping in sensor validation: a comparison of the filter and wrapper approaches", Journal of Risk and Reliability (JRR), 2008, Proc. IMechE, Vol. 222, Part O, pp. 189-206.

[13] Perrone M.P., and Cooper L.N., When Networks Disagree: Ensemble Methods for Hybrid Neural Networks. In R.J. Mammone (ed.), Neural Networks for Speech and Image Processing. New York: Chapman & Hall; 1993.

[14] A. Krogh and J. Vedelsby, "Neural network ensembles, cross-validation and active learning", in: G. Tesauro, D. S. Touretzky and T. K. Loen, 1995, Advances in newel information processing systems, Vol. 7, pp. 231-238, MIT press, Cambridge, MA, USA.

[15] A. J. C. Sharkey, "On combining artificial neural nets", 1996, Connection Science, Vol. 8(3), pp. 299-314.

[16] G. Gola, E. Zio, P. Baraldi, D. Roverso, M. Hoffmann, "Signal Grouping for Sensor Validation: a Multi-Objective Genetic Algorithm Approach", HWR-852, OECD Halden Reactor Project, Feb. 2007.

[17] P. Baraldi, E. Zio, G. Gola, D. Roverso, M. Hoffmann, Reconstruction of faulty signals by an ensemble of principal component analysis models optimized by a multi-objective genetic algorithm, Proceedings of FLINS, Madrid, Spain, Sept. 2008.

[18] G. Gola, E. Zio, P. Baraldi, D. Roverso, M. Hoffmann, Reconstructing signals for sensor validation by a GA-optimized ensemble of PCA models, HWR-894, OECD Halden Reactor Project, Apr. 2008.

[19] N. Rooney, D. Patterson, S. Anand, A. Tsymbal, Dynamic Integration of Regression Models, 2204, Lecture notes in computer science, MCS 2004: Multiple Classifier Systems, 5th International workshop, Cagliari, Italy, Vol. 3077, pp. 164-173, Springer, Berlin, Germany.

[20] G. Ratsch, A. Demiriz, K. P. Bennett, Sparse Regression Ensembles in Infinite and Finite Hypothesis Spaces, 2002, Machine Learning, Vol. 48, pp. 189-218.

[21] G. Brown, J. L. Wyatt, P. Tino, Managing Diversity in Regression Ensembles, 2005, Journal of Machine Learning Research, Vol. 6, pp. 1621-1650.

[22] Y. Yu, Z.-H. Zhou, K. Ming Ting, Cocktail Ensemble for Regression, 2007, Proc. of 7th IEEE International Conference on Data Mining, pp.721-726.

[23] L. Breiman, Bagging predictors, 1996, Machine Learning, Vol. 24, pp. 123-140.

[24] R. Polikar, Ensemble based systems in decision making, 2006, IEEE Circuits and Systems Magazine, Vol. 6(3), pp. 21-45.

[25] A. Tsymbal, S. Puuronen, I. Skrypnyk, Ensemble feature selection with dynamic integration of classifiers, in: Int. ICSC Congress on Computational Intelligence Methods and Applications CIMA' 2001, Bangor, Wales, UK, 2001, pp. 558-564.

[26] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in search strategy for ensemble feature selection, 2005, Information Fusion, Vol. 6, pp. 83-98.

[27] R. Bryll, R. Gutierrez-Osuna, F. Quek, Attribute bagging: improving accuracy of classifiers ensembles by using random feature subsets, 2003, Pattern Recognition, Vol. 36, pp. 1291-1302.

[28] I.T. Jolliffe, "Principal Component Analysis", Springer Eds., 2002.

[29] K.I. Diamantaras, S.Y. Kung, "Principal component neural networks: theory and applications", John Wiley & Sons, Inc. New York, NY, USA, 1996.

[30] B. Scholkopf, A. Smola and K.R. Muller, "Kernel principal component analysis", Advances in Kernel Methods-Support Vector Learning, 1999.

[31] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction", IEEE Transactions on Automatic Control, Vol. 26, Issue 1, Feb. 1981.

[32] P. P. Bonissone, F. Xue, R. Subbu, Fast Meta-Models for Local Fusion of Multiple Predictive Models, Applied Soft Computing, Article in Press, available on-line from March 20, 2009.

[33] L. I. Kuncheva, "Classifier Ensembles for Changing Environment", Multiple Classifier Systems, in: "Lecture Notes on Computer Science, Vol. 3077, 2004, Springer Berlin / Heidelberg Eds.

[34] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, "On Combining Classifiers", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, Vol. 20 (3), pp. 226-239

[35] J.H. Holland, "Adaptation in natural and artificial systems: an introductory analysis with application to biology", Control and Artificial Intelligence. MIT Press, 4-th edition, 1975.

[36] D.E. Goldberg, "Genetic algorithms in search, optimization, and machine learning", Addison-Wesley Publ. Co., 1989.

[37] L. Chambers, "Practical handbook of genetic algorithms: applications Vol. I; new frontiers Vol. II", CRC Press, 1995.

[38] Y. Sawaragy, H. Nakayama, T. Tanino, "Theory of multiobjective optimization", Academic Press, Orlando, Florida, 1985.

[39] M. Marseguerra, Lecture notes on Principal Components Analysis (PCA), Polytechnic of Milan.

Appendix. Principal Component Analysis (PCA) for signal validation and reconstruction

In this Appendix, we briefly sketch the procedural steps of Principal Component Analysis (PCA) as presented in [39].

The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated and ordered so that the first *few* retain the most of the variation present in *all* of the original variables [28].

In this view, let $\{f_i(t), t=1,2,...,N, i=1,2,...,m\}$ be a set of N observations in an m-dimensional space $\Re^m$. The purpose of the PCA is to identify a λ-dimensional ($\lambda < m$) subspace $\Re^\lambda \subset \Re^m$ in which the most of the data set variation is retained and the least information is lost.

From a mathematical point of view, let $\mathbf{X} \equiv (N, m)$ be the data set matrix whose rows $\mathbf{f}_t \equiv (1, m), t=1,2,...,N$, are the patterns of the m observations, i.e. the m signal values at the time instant t, viz.:

$$\mathbf{X} = \begin{pmatrix} f_{11}\ f_{12} \cdots f_{1m} \\ f_{21}\ f_{22} \cdots f_{2m} \\ \cdots \\ f_{N1}\ f_{N2} \cdots f_{Nm} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \cdots \\ \mathbf{f}_N \end{pmatrix} \equiv (N, m) \tag{A1}$$

so that $X_{ti} = f_{ti}$, $t = 1,2,...,N, i = 1,2,...,m$, is the i-th component of $\mathbf{f}_t$ in the original basis.

Let $\mathbf{P} \equiv (m,m) \in \Re^m$ be a matrix constituted by m orthonormal column vectors $\mathbf{p}_i \equiv (m,1), i = 1,2,...,m$ built from the data set $\mathbf{X}$ according to an optimality criterion to be defined later and representing an orthonormal basis for the data set:

$$\mathbf{P} = \begin{pmatrix} p_{11}\ p_{12} \cdots p_{1m} \\ p_{21}\ p_{22} \cdots p_{2m} \\ \cdots \\ p_{m1}\ p_{m2} \cdots p_{mm} \end{pmatrix} = (\mathbf{p}_1\ \mathbf{p}_2 \ldots \mathbf{p}_m) \equiv (m,m) \tag{A2}$$

so that $P_{ij} = p_{ij}, i = 1,2,...,m,\ j = 1,2,...,m$ is the j-th component of $\mathbf{p}_i$ in the original basis and, for the orthonormality of the basis vectors, $\mathbf{p}_i^T \cdot \mathbf{p}_j = \delta_{ij}$ or $\mathbf{P}^T \cdot \mathbf{P} = \mathbf{I}_m$, where $\mathbf{I}_m$ is the unit matrix of order m.

In the orthonormal basis, let u_{ti} be the component of $\mathbf{f}_t$ along $\mathbf{p}_i^T$, so that

$$f_{tj} = \sum_{i=1}^{m} u_{ti} \cdot p_{ij} \tag{A3}$$

and

$$\mathbf{f}_t = \sum_{i=1}^{m} u_{ti} \cdot \mathbf{p}_i^T = (u_{t1}\ u_{t2} \ldots u_{tm}) \begin{pmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \\ \cdots \\ \mathbf{p}_m^T \end{pmatrix} = \mathbf{u}_t \cdot \mathbf{P}^T \tag{A4}$$

where $\mathbf{u}_t \equiv (1,m), t = 1,2,...,N$, is the t-th m-dimensional pattern constituted by the m signal values at time t in the orthonormal basis.

Right multiplying Eq. (A4) by $\mathbf{P}$ yields,

$$\mathbf{u}_t = \mathbf{f}_t \cdot \mathbf{P} \tag{A5}$$

and in matrix form

$$\mathbf{U} = \mathbf{X} \cdot \mathbf{P} \tag{A6}$$

where $\mathbf{U} \equiv (N,m)$ is the matrix whose N rows $\mathbf{u}_t \equiv (1,m)$ are the coordinates of $\mathbf{f}_t$, $t = 1,2,...,N$, in the orthonormal basis.

The data set has now two representations: when intended in the original basis, the t-th pattern is the vector $\mathbf{f}_t$ with components f_{ti}; when intended in the orthonormal $\mathbf{P}$ basis, the same t-th pattern is a vector $\mathbf{u}_t$ with component u_{ti} along $\mathbf{p}_i$. In this view, once the orthonormal $\mathbf{P}$ basis has been fixed, Eq. (A5) provides $\mathbf{u}_t$ as a function of $\mathbf{f}_t$. To get the inverse relation, we right multiply by $\mathbf{P}^T$ and we obtain the data set $\mathbf{X}$ in the original basis, viz.:

$$\mathbf{X} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \cdots \\ \mathbf{f}_N \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \cdots \\ \mathbf{u}_N \end{pmatrix} \cdot \mathbf{P}^T = \mathbf{U} \cdot \mathbf{P}^T \equiv (N,m)(m,m) \tag{A7}$$

In this view, Eqs. (A6) and (A7) represent the transformation laws of the observation patterns $\mathbf{X}$ between the original reference system and the orthonormal basis. Notice that up to this point the equations are exact and the data values are transformed in both senses without any loss of information.

The PCA approximation consists in mapping the observation vectors $\mathbf{f}_t$ in a subspace $\Re^{\lambda} \subset \Re^m$ identified by $\lambda < m$ vectors chosen according to a criterion explained later among the $\mathbf{p}_i, i = 1,2,...,m$.

Without loss of generality, assume that the basis vectors are ordered in such a way that the selected λ vectors are the first ones in $\mathbf{P}$, i.e. $(\mathbf{p}_1\ \mathbf{p}_2\ ...\mathbf{p}_{\lambda})$. Correspondingly, the matrices $\mathbf{P}$ and $\mathbf{U}$ are partitioned as follows:

$$\mathbf{P} = (\mathbf{P}_{\lambda}\ \mathbf{P}_{m-\lambda}) \text{ and } \mathbf{U} = (\mathbf{U}_{\lambda}\ \mathbf{U}_{m-\lambda})$$

where $\mathbf{P}_{\lambda} \equiv (m, \lambda)$ and $\mathbf{U}_{\lambda} \equiv (N, \lambda)$ are the submatrices constituted by the first λ columns of $\mathbf{P}$ and $\mathbf{U}$, respectively, and $\mathbf{P}_{m-\lambda} \equiv (m, m-\lambda)$ and $\mathbf{U}_{m-\lambda} \equiv (N, m-\lambda)$ are the submatrices constituted by the last $m-\lambda$ columns of $\mathbf{P}$ and $\mathbf{U}$, respectively. The column vectors in $\mathbf{P}_{\lambda}$ and $\mathbf{P}_{m-\lambda}$ constitute the bases of the two mutually orthogonal subspaces $\Re^{\lambda}$ and $\Re^{m-\lambda}$ in which $\Re^m$ has been divided. In terms of the above submatrices, Eq. (A7) can be rewritten as:

$$\mathbf{X} = (\mathbf{U}_{\lambda}\ \mathbf{U}_{m-\lambda}) \begin{pmatrix} \mathbf{P}_{\lambda}^T \\ \mathbf{P}_{m-\lambda}^T \end{pmatrix} = \mathbf{U}_{\lambda} \cdot \mathbf{P}_{\lambda}^T + \mathbf{U}_{m-\lambda} \cdot \mathbf{P}_{m-\lambda}^T = \tilde{\mathbf{X}} + \mathbf{U}_{m-\lambda} \cdot \mathbf{P}_{m-\lambda}^T \tag{A8}$$

where we define $\tilde{\mathbf{X}}$ as:

$$\tilde{\mathbf{X}} = \mathbf{U}_{\lambda} \cdot \mathbf{P}_{\lambda}^T = \begin{pmatrix} u_{11}\ u_{12}\ ...u_{1\lambda} \\ u_{21}\ u_{22}\ ...u_{2\lambda} \\ ... \\ u_{N1}\ u_{N2}\ ...u_{N\lambda} \end{pmatrix} \begin{pmatrix} \mathbf{P}_1^T \\ \mathbf{P}_2^T \\ ... \\ \mathbf{P}_{\lambda}^T \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{f}}_1 \\ \tilde{\mathbf{f}}_2 \\ ... \\ \tilde{\mathbf{f}}_N \end{pmatrix} \equiv (N, m) \tag{A9}$$

The t-th row of $\tilde{\mathbf{X}}$, namely $\tilde{\mathbf{f}}_t$, is the orthonormal projection of $\mathbf{f}_t$ onto $\Re^{\lambda}$ and then it may be expressed by as a linear combination of the vectors of the $\mathbf{P}_{\lambda}$ basis, viz.:

$$\tilde{\mathbf{f}}_t = (u_{t1}\ u_{t2}\ ...u_{t\lambda}) \begin{pmatrix} \mathbf{P}_1^T \\ \mathbf{P}_2^T \\ ... \\ \mathbf{P}_{\lambda}^T \end{pmatrix} = \sum_{i=1}^{\lambda} u_{ti} \cdot \mathbf{P}_i^T \tag{A10}$$

and

$$\tilde{f}_{tj} = \sum_{i=1}^{\lambda} u_{ti} \cdot p_{ij} \tag{A11}$$

is the j-th component, $j = 1,2,...,m$, of $\tilde{\mathbf{f}}_t$ in the original basis in $\Re^m$ expressed through the components u_{ti} and p_{ij}, $i = 1,2,...,\lambda$ of $\mathbf{u}_t$ and $\mathbf{p}_j$ in $\Re^{\lambda}$.

If all the information about the data set $\mathbf{X}$ essentially lies in a λ-dimensional space $\Re^{\lambda}$ (apart from small components in $\Re^{m-\lambda}$ given by $\mathbf{U}_{m-\lambda} \cdot \mathbf{P}_{m-\lambda}^T$ as stated in Eq. A8), then the data analysis can be performed in $\Re^{\lambda}$ reducing the dimension of the data set to handle by a factor λ / m. To this aim, each observation vector $\mathbf{f}_t \in \Re^m$ is approximated by its orthonormal projection $\tilde{\mathbf{f}}_t \in \Re^{\lambda}$ plus a residual vector in $\Re^{m-\lambda}$ which is postulated to be independent of t, viz.,

$$\tilde{\mathbf{f}}_t^{appx} = \tilde{\mathbf{f}}_t + \sum_{i=\lambda+1}^{m} b_i \cdot \mathbf{p}_i^T \tag{A12}$$

The best residual vector is that one which, on the average, minimizes the absolute value of the square error between the real $\{\mathbf{f}_t\}$ and approximated $\{\tilde{\mathbf{f}}_t^{appx}\}$ data patterns, i.e.:

$$E = \frac{1}{2}\sum_{t=1}^{N}\left\|\mathbf{f}_t - \tilde{\mathbf{f}}_t^{appx}\right\|^2 \tag{A13}$$

By combining Eqs. (A4), (A10) and (A12), the error between the two vectors can be written as:

$$\mathbf{f}_t - \tilde{\mathbf{f}}_t^{appx} = \sum_{i=\lambda+1}^{m}\left(u_{ti} - b_i\right)\cdot\mathbf{p}_i^T \tag{A14}$$

and

$$\left\|\mathbf{f}_t - \tilde{\mathbf{f}}_t^{appx}\right\|^2 = \left(\mathbf{f}_t - \tilde{\mathbf{f}}_t^{appx}\right)\left(\mathbf{f}_t - \tilde{\mathbf{f}}_t^{appx}\right)^T = \sum_{i=\lambda+1}^{m}\left(u_{ti} - b_i\right)\cdot\mathbf{p}_i^T\sum_{j=\lambda+1}^{m}\left(u_{tj} - b_j\right)\cdot\mathbf{p}_j \tag{A15}$$

The expression for the error becomes then,

$$E = \frac{1}{2}\sum_{t=1}^{N}\sum_{i=\lambda+1}^{m}\left(u_{ti} - b_i\right)^2 \tag{A16}$$

The best constants are those that minimize the error and are determined by the conditions

$$\frac{\partial E}{\partial b_i} = -\sum_{t=1}^{N}\left(u_{ti} - b_i\right) = 0 \tag{A17}$$

Since the constants b_i, $i = \lambda+1,...,m$, do not depend on t, using Eq. (A4) we can write,

$$b_i = \frac{1}{N}\sum_{t=1}^{N}u_{ti} = \left(\frac{1}{N}\sum_{t=1}^{N}\mathbf{f}_t\right)\cdot\mathbf{p}_i = \bar{\mathbf{f}}\cdot\mathbf{p}_i \tag{A18}$$

where the vector $\bar{\mathbf{f}}$ is the arithmetic average of the observation vectors, i.e. the average value of the signals. In particular, the i-th component of $\bar{\mathbf{f}}$ is the arithmetic average of the i-th column of $\mathbf{X}$.

Then, from Eq. (A12), the expression for the PCA approximation of the data pattern $\mathbf{f}_t$ is given by:

$$\tilde{\mathbf{f}}_t^{appx} = \tilde{\mathbf{f}}_t + \bar{\mathbf{f}}\sum_{i=\lambda+1}^{m}\mathbf{p}_i\cdot\mathbf{p}_i^T \tag{A19}$$

or in matrix form,

$$\tilde{\mathbf{X}}^{appx} = \tilde{\mathbf{X}} + \bar{\mathbf{X}}\cdot\mathbf{P}_{m-\lambda}\cdot\mathbf{P}_{m-\lambda}^T \tag{A20}$$

Coming to the problem of model learning, the Principal Components Analysis exploits the information in the m-dimensional data set to generate an orthonormal basis $\mathbf{P}$ in $\Re^m$ constituted by m eigenvectors. These represent the result of the learning phase of the PCA model.

In this respect, let $\mathbf{V}$ be the covariance matrix of the data set. The problem to tackle at this point is how to choose an orthonormal basis $\mathbf{P}$ in $\Re^m$ and how to select among the m columns $\mathbf{p}_i$ of $\mathbf{P}$ the λ vectors which constitute the basis of $\Re^\lambda$. As demonstrated and explained in details in [28, 29, 39], by substituting Eqs. (A4) and (A18) into Eq. (A16), we can write the minimum error corresponding to the coefficients b_i, $i = \lambda+1,...,m$, as:

$$E_{\min} = \frac{1}{2}Tr\left[\mathbf{P}_{m-\lambda}^T\cdot\mathbf{V}\cdot\mathbf{P}_{m-\lambda}\right] \tag{A21}$$

where $\mathbf{V}$ represents the covariance matrix of $\mathbf{X}$ eventually positive definite (so that its eigenvalues are real and positive) [28] and can be written as,

$$\mathbf{V} = \sum_{t=1}^{N} \left(\mathbf{f}_t - \bar{\mathbf{f}}\right)^T \left(\mathbf{f}_t - \bar{\mathbf{f}}\right) = \left(\mathbf{X} - \overline{\mathbf{X}}\right)^T \left(\mathbf{X} - \overline{\mathbf{X}}\right) \equiv (m, N)(N, m) \tag{A22}$$

In order to find the λ vectors which will constitute $\mathbf{P}_\lambda$, let us minimize $E_{\min}$ with respect to $\mathbf{P}_{m-\lambda}$ by resorting to the Lagrange multiplier approach [28]. The purpose is to find those $m-\lambda$ vectors which minimize $E_{\min}$ subject to the constraint of being orthonormal. The Lagrange function in terms of the submatrix $\mathbf{P}_{m-\lambda}$ can be written as:

$$L = \frac{1}{2} Tr\left[\mathbf{P}_{m-\lambda}^T \cdot \mathbf{V} \cdot \mathbf{P}_{m-\lambda}\right] - \frac{1}{2} Tr\left[\mathbf{\Lambda}_{m-\lambda}\left(\mathbf{P}_{m-\lambda}^T \cdot \mathbf{P}_{m-\lambda} - \mathbf{I}_{m-\lambda}\right)\right] \tag{A23}$$

where $\mathbf{\Lambda}_{m-\lambda} \equiv (m-\lambda)(m-\lambda)$ is the matrix of the Lagrange coefficients, namely Λ_{ij}, and $\mathbf{I}_{m-\lambda}$ is the unit matrix of order $m-\lambda$.

Differentiating L with respect to $\mathbf{P}_{m-\lambda}$ and setting to zero the result we obtain [28]:

$$\mathbf{V} \cdot \mathbf{P}_{m-\lambda} = \mathbf{P}_{m-\lambda} \cdot \mathbf{\Lambda}_{m-\lambda} \tag{A24}$$

One solution to this equation is to choose $\mathbf{\Lambda}_{m-\lambda}$ to be diagonal, i.e. $\left(\mathbf{\Lambda}_{m-\lambda}\right)_{ij} = \mathbf{\Lambda}_i \cdot \delta_{ij}$, so that the columns of $\mathbf{P}_{m-\lambda}$ are the eigenvectors of $\mathbf{V}$ corresponding to the eigenvalues $\Lambda_i, i = \lambda+1,...,m$. Notice that since the eigenvalues have been supposed simple, the eigenvectors are orthogonal and may be normalized.

By substituting (A24) into (A23) we obtain that the required minimum of the Lagrangian is

$$\hat{L} = \hat{E}_{\min} = \frac{1}{2} Tr\left[\mathbf{\Lambda}_{m-\lambda}\right] = \frac{1}{2} \sum_{i=\lambda+1}^{m} \Lambda_i \tag{A25}$$

In principle, any set of $m-\lambda$ eigenvectors can constitute the orthonormal basis in $\Re^{m-\lambda}$, but from Eq. (A25) it appears that the best choice is to select the smallest $m-\lambda$ eigenvalues among the m possible eigenvalues in $\mathbf{V}$. To this aim, we rank the m eigenvalues in decreasing order so that

$$\Lambda_1 > \Lambda_2 > ... > \Lambda_m$$

The eigenvectors are correspondingly ranked and we choose for the basis $\mathbf{P}_\lambda$ of $\Re^\lambda$ the first λ eigenvectors and for the basis $\mathbf{P}_{m-\lambda}$ of $\Re^{m-\lambda}$ the remaining $m-\lambda$ ones. The amount of information lost by considering $\tilde{\mathbf{X}}^{appx}$ instead of $\mathbf{X}$ may be quantified for the individual observations by the differences $\mathbf{f}_t - \tilde{\mathbf{f}}_t^{appx}$ or globally by the fraction of neglected eigenvalues, namely $2E_{\min} \Big/ \sum_{i=1}^{m} \Lambda_i$.

Finally, coming to the problem of signal validation and reconstruction by means of a PCA-based model, in order to simplify the calculations, the time trends of the signals have been previously normalized so that their mean is zero and their standard deviation equals 1. This allows to skip the computation of the residuals, since $\bar{\mathbf{f}} = 0$ and, according to Eq. (A19), $\tilde{\mathbf{f}}_t^{appx} = \tilde{\mathbf{f}}_t$.

Furthermore, coming to the problem of using the PCA as a signal reconstruction model, for each group k constituted by m_k signals, the eigenvectors constituting the orthonormal basis $\mathbf{P}_\lambda$, have been obtained by Eq. (A24) from the covariance matrix $\mathbf{V}$ of the pairwise signal correlations between the m_k signals in the group.